Incubation time of AIDS

In predictions of the development of AIDS, the length of the incubation time is one of the key things we have to estimate.

Problems:

1. We usually do not know the seroconversion time of a person who has AIDS.

2. We often do not have precise information about the onset of AIDS for a person who is sero-positive.

The sero-conversion time is usually *interval censored*
The time of onset of AIDS is often *right censored*.

Interval censoring, case 2

$X_1, \ldots, X_n$ random sample generated by a distribution function F.
$(T_1, U_1), \ldots, (T_n, U_n)$ random sample generated by a distribution function H

X_i's independent of (T_i, U_i)'s; $X_i \geq 0$
(T_i, U_i)'s: observation times; $U_i > T_i \geq 0$

X_i's are only indirectly observable via:

$$\Delta_i^{(1)} = \begin{cases} 1, & X_i \leq T_i \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\Delta_i^{(2)} = \begin{cases} 1, & T_i < X_i \leq U_i \\ 0, & \text{otherwise.} \end{cases}$$

Maximum likelihood estimator of F: a distribution function maximizing:

$$(1) \quad \prod_{i=1}^{n} F(t_i)^{\delta_i^{(1)}} (F(u_i) - F(t_i))^{\delta_i^{(2)}} \cdot (1 - F(u_i))^{1-\delta_i^{(1)}-\delta_i^{(2)}} h(t_i, u_i),$$

as a function of F.

Computation of the (NP)MLE of the distribution function boils down to solving the (finite-dimensional) optimization problem:

Maximize (1) under the restrictions:

$$0 \leq F(t_i), F(u_i) \leq 1,$$

and, for $x, y \in \{z : z = t_i \text{ or } z = u_i,\ i = 1, \ldots, n\}$ with $x < y$:

$$F(x) \leq F(y),$$

Some theory

Let $\widehat{F}_n$ be an MLE of the distribution function F. Then $\widehat{F}_n(t)$ is a *consistent* estimator of $F(t)$ at a point t where one of the marginal distributions of the pairs of observation times (T_i, U_i) have a positive density.
This means

$$\widehat{F}_n(t) \to F(t), \text{ almost surely, as } n \to \infty.$$

Moreover, if $P\{U_i - T_i > \epsilon\} = 1$, for some $\epsilon > 0$, then (Groeneboom (St-Flour, 1996)):

$$n^{1/3}\left\{\widehat{F}_n(t) - F(t)\right\} \to cZ, \text{ as } n \to \infty,$$

in distribution, as $n \to \infty$, where Z is the location of the maximum of the process

$$\{W(x) - x^2 : x \in I\!R\}$$

and W is two-sided (standard) Brownian motion. The positive constant c is a function of F and H (and can be computed).

Conjecture (10 years old): if the density of (T_i, U_i) is positive on the diagonal, then

$$\{n \log n\}^{1/3} \left\{ \widehat{F}_n(t) - F(t) \right\} \to cZ, \text{ as } n \to \infty,$$

in distribution, as $n \to \infty$.

On the other hand, "smooth functionals" of the model, like the mean, can be estimated at the usual $\sqrt{n}$-rate, and have a normal limiting distribution:

$$n^{1/2} \left\{ \int x \, d\widehat{F}_n(x) - \int x \, dF(x) \right\} \to U, \text{ as } n \to \infty,$$

in distribution, as $n \to \infty$, where U has a normal distribution with mean zero and a variance σ^2 that can be found by solving a (non-standard) integral equation (Geskus and Groeneboom (1999)).

Right censoring

Model:
$X_1, \ldots, X_n$ random sample generated by a distribution function F.
$C_1, \ldots, C_n$ random sample of *censoring times* generated by a distribution function G.

X_i's independent of C_i's; $X_i \geq 0$

X_i's are either directly or indirectly observable via the random variables (Y_i, Δ_i), given by

$$(Y_i, \Delta_i) = (\min\{X_i, C_i\}, \Delta_i)$$

$$= \begin{cases} (X_i, 1), & X_i \leq C_i \\ (C_i, 0), , & \text{otherwise.} \end{cases}$$

Likelihood for the distribution F:

$$\prod_{i=1}^{n} f(y_i)^{\delta_i} \{1 - F(y_i)\}^{1-\delta_i}. \tag{2}$$

where f is the probability density, corresponding to F.

If we restrict the set of distribution functions F to the purely discrete distribution, with mass at the points y_i (and an extra point at ∞ for a "fictitious supersurvivor"), then the MLE $\widehat{F}_n$ of F is given by the *Kaplan-Meier* estimator:

$$\widehat{F}_n(t) = \prod_{i:y_i \leq t} \frac{\#\{j : y_j \geq y_i\} - \delta_i}{\#\{j : y_j \geq y_i\}}.$$

Differences between right censoring and interval censoring

1. The Kaplan-Meier estimator has an explicit expression as a function of the sample. No iterative algorithm is needed to compute it, in contrast with the MLE in the case of interval censoring.

2. Asymptotic theory for the Kaplan-Meier estimator is relatively easy: at interior points t of the support we have:

$$\sqrt{n}\left\{\widehat{F}_n(t) - F(t)\right\} \to N\left(0, \sigma^2\right),$$

where $\to$ means convergence in distribution, and $N\left(0, \sigma^2\right)$ denotes a normal distribution with a variance that can be explicitly expressed as function of F and the censoring distribution.

3. The *process* $\left\{\sqrt{n}\left\{\widehat{F}_n(t) - F_0(t)\right\} : t \geq 0\right\}$ converges on compact intervals of the support of the distribution of F to a Gaussian process, if $\widehat{F}_n$ is the Kaplan-Meier estimator. No process convergence if $\widehat{F}_n$ is the MLE for the interval censoring model!

4. Martingales can be used for deriving asymptotic theory for the Kaplan-Meier estimator. Martingales are useless for the interval censoring model!

How can one estimate the distribution of the incubation time?

One possible method:
Estimate the joint distribution function of seroconversion time and time of onset of AIDS. Say this gives the estimate $\widehat{H}_n$. Then estimate the incubation time distribution function by

$$\widehat{F}_n(t) = \int_{(x,y):y-x\leq t} d\widehat{H}_n(x,y).$$

Other method: (DeGruttola and Lagakos (1989), Geskus (2000): *Double censoring:*
Assume:

(i) Seroconversion time and incubation time are independent random varianbles.

(ii) Conditionally on serostatus at entry, the observation times are independent of both seroconversion time and incubation time.

Model for this approach:
F: seroconversion df with density f
G: incubation time df with density g
v_i^n: observation time where ith person was still seronegative
v_i^p: observation time where the person was seropositive.
(z_i, δ_i): an observation time and an indicator for the ith person, where
either:
z_i is the time of onset of AIDS, and $\delta_i = 1$,
or:
z_i is a right-censored observation time for the onset of AIDS and $\delta_i = 0$.
Now maximize

$$\sum_{i=1}^{n} \Big\{ \delta_i \log \int_{v_i^n}^{v_i^p} g(z_i - x)\, dF(x) + (1 - \delta_i) \log \int_{v_i^n}^{v_i^p} \{1 - G(z_i - s)\}\, dF(x) \Big\},$$

as a function of F and G.

Problems:

(i) Assumption of convolution structure

(ii) Criterion function is not concave in the relevant parameters

(iii) We can make the criterion function arbitrarily large by taking, if $\delta_i = 1$:
$v_i \in (v_i^n, v_i^p)$ and

$$g(z_i - s) = \frac{c}{\sqrt{v_i - s}},\ f(s) = \frac{c}{\sqrt{v_i - s}},$$

if $s \in [v_i - \epsilon, v_i)$, and zero elsewhere, where one can take $\epsilon > 0$ arbitrarily small.

"Solution" of difficulty (iii): take a grid which is "not too fine". But what is "not too fine"?

Bivariate current status data
(simplest case of bivariate interval censoring)

Model: $X = (X, Y)$ is a random variable with df H.
We *observe*: $(U, V, \Delta_1, \Delta_2)$, where $T = (U, V)$ is a pair of observation times with df F, independent of X, and where

$$\Delta_1 = 1_{\{X \leq U\}},\ \Delta_2 = 1_{\{Y \leq V\}}.$$

Example: (Hughes, Richardson (2000)): Vertical transmission of HIV.

MLE $\hat{H}_n$ of H:

$$\begin{aligned} \text{argmax}_H \int \Big\{ & \delta_1 \delta_2 \log H(u, v) \\ & + \delta_1 \overline{\delta}_2 \log H^{(2)}(u, v) \\ & + \overline{\delta}_1 \delta_2 \log H^{(3)}(u, v) \\ & + \overline{\delta}_1 \overline{\delta}_2 \log H^{(4)}(u, v) \Big\}\, d\,\mathbb{P}_n \end{aligned}$$

where $H^{(2)}(u, v) = H(u, \infty) - H(u, v)$, etc. (the 4 quadrants).

How can one compute the MLE?

1. EM?
   Did not converge in two days on Sun workstation for sample size 500 (Song (2000)) with an accuracy 10^{-5}

2. Vertex exchange algorithm?
   (Gentleman, Vandal (1999))
   Did not converge *at all* for sample size 200 (went into a loop) (idem).
   Same accuracy criterion as for EM

3. Primal-dual interior point method?
   Took 45 minutes for sample size 500 at accuracy 10^{-10} (idem).

Can asymptotic theory help?

Conjecture: the number of points of support of the MLE will generally be of order $n^{1/3}$.

Suggests a *vertex direction algorithm*, systematically looking for the points of support, with corresponding masses, starting from scratch with zero points.

EM, vertex exchange algorithm and primal-dual interior point method all work on a grid of order n^2 points and try to "work back" to the order $n^{1/3}$ points of support.

C program finding the exact solution (in 10 decimals) and the points of support, using a vertex direction algorithm is now available.

Some theory (Song (2001)):
Global convergence rate of MLE: $n^{-3/10}$
(Probably not sharp).

Local minimax lower bound of order $n^{-1/3}$.
This holds for any fixed dimension!

Proof technique:
Make a perturbation of the density on a square with sides of order $n^{-1/6}$ with midpoint (u_0, v_0) (a point in the interior of the support of the distribution). Then use an inequality of the type:

$$\inf_{T_n} \max_{g_1, g_2} \{E_{g_1}|T_n - Tg_1|, E_{g_2}|T_n - Tg_2|\}$$
$$\geq \tfrac{1}{4}|Tg_1 - Tg_2| \left\{1 - H^2(g_1, g_2)\right\}^{2n}.$$